치의학박사 학위논문

Comparison among bone age assessment methods and development of a Fishman-based skeletal maturity determination system using deep learning in contemporary Korean children and adolescents

현대 한국 소아 및 청소년에 있어 골령 평가법 비교 및 Fishman법 기반의 골격 성숙도 측정 딥러닝 시스템 개발

2020년 8월

서울대학교 대학원

치의과학과 영상치의학 전공

신 난 영

Comparison among bone age assessment methods and development of a Fishman-based skeletal maturity determination system using deep learning in contemporary Korean children and adolescents

지도교수 허 민 석
이 논문을 치의학박사학위논문으로 제출함

2020년 5월

서울대학교 대학원
치의과학과 영상치의학 전공
신 난 영

신난영의 치의학박사학위논문을 인준함

2020년 7월

위 원 장 _____ (인)
부위원장 _____ (인)
위 원 _____ (인)
위 원 _____ (인)
위 원 _____ (인)

ABSTRACT

Comparison among bone age assessment methods and development of a Fishman-based skeletal maturity determination system using deep learning in contemporary Korean children and adolescents

Nan-Young Shin, DDS

Department of Oral and Maxillofacial Radiology,
Graduate School, Seoul National University
(Directed by Prof. Min-Suk Heo, DDS, MSD, PhD)

Purpose

Greulich-Pyle (GP), Tanner-Whitehouse 3 (TW3), and Fishman methods are typically employed for bone age assessment from hand-wrist radiographs. This study aimed to compare the Fishman method with the GP and TW3 methods and investigate the reliability of Fishman's skeletal maturity indicators (SMIs) for contemporary healthy Korean children and adolescents, and to develop a new fully-automated SMI-based skeletal maturity determination system using deep neural networks and evaluate the accuracy of the system.

Materials and Methods

The left hand-wrist radiographs of 1,617 subjects (706 males and 911 females; 6–17 years of age) taken in 2012–2017 were selected. Bone ages were calculated using the GP, TW3, and Fishman methods, and compared with chronological ages using paired t-test and correlation analysis. For developing a fully-automated deep learning system for skeletal maturity determination using the Fishman method, two skeletal maturity determination systems were developed and their accuracies were compared. A system was trained with an SMI-labeled dataset, and another one was trained with a dataset that was not only labeled with SMIs but was additionally labeled considering the region of interest (ROI) extraction and skeletal maturity determination for each ROI. Two oral and maxillofacial radiologists established a reference standard for the SMIs.

Results

The bone ages significantly differed with the chronological ages in the whole group and gender subgroups for all three methods except in the male group for the TW3 method. However, a high degree of correlation was observed between the chronological ages and the bone ages when evaluated by all the methods. For the skeletal maturity determination system that was trained using the dataset labeled with only SMIs, the mean absolute error (MAE) was 0.88 and the within-1 concordance rate was 73.1 %. Conversely, the system consisting of ROI extraction, ROI-based skeletal maturity determination, and final SMI prediction showed much better outcomes; the MAE was 0.34 and the within-1 concordance rate was 93.7 %.

Conclusions

In this study, Fishman's SMI was confirmed as a reliable index for the determination of skeletal maturity from hand-wrist radiographs. A developed deep learning system automated the entire process consisting of ROI extraction, skeletal maturity determination for each ROI, and final SMI prediction. The system's accuracy in predicting skeletal

maturity was outstanding. Thus, the system presented in this study can be applied effectively to determine the skeletal maturity of contemporary Korean children and adolescents.

Keywords: Skeletal Maturity Indicator, Hand-Wrist Radiograph, Fishman Method, Deep Learning

Student number: 2017-35685

Contents

I. Introduction 1

II. Materials and Methods 8

III. Results 21

IV. Discussion 33

V. Conclusion 38

VI. References 39

Abstract(Korean) 43

List of tables

Table 1. Approximate chronological ages and percentage of growth completed corresponding to skeletal maturity indicators (SMIs)
……..……...5
Table 2. Sample distribution according to age and gender
.…….……..……18
Table 3. Skeletal maturation stages for each region of interest (ROI) defined in the study
….…….……..……..……19
Table 4. Combination of skeletal maturation stages of regions of interests (ROIs) for each skeletal maturity indicator (SMI)
…….……..……..……20
Table 5. Means and standard deviations of chronological ages (CAs) and bone ages (BAs) estimated by GP method in each age group (unit: year)
..…….……..……24
Table 6. Means and standard deviations of chronological ages (CAs) and bone ages (BAs) estimated by TW3 method in each age group (unit: year)
….…….……..……25
Table 7. Means and standard deviations of chronological ages (CAs) and bone ages (BAs) estimated by Fishman method in each age group (unit: year)
….…….……..……26
Table 8. Correlation coefficients between chronological age and bone age (GP method and TW3 method) or skeletal maturity indicator (Fishman method) measured by each method
….…….……..……27
Table 9. Inter- and intra-observer reliability for TW3 method and Fishman method
….…….……..……27
Table 10. Sample distribution with reference to skeletal maturity indicator (SMI) and comparison between the matched mean chronological ages (CAs) in this study and that in Fishman's study (unit: year)
…….……..……28

Table 11. Comparison between the mean age intervals based on skeletal maturity indicator (SMI) in this study and that in Fishman's study (unit: year)
……..……..……..……29
Table 12. Test data distribution and skeletal maturity indicator (SMI) prediction accuracy of the system trained with only SMIs
……..……..30
Table 13. Accuracy of the system trained using each region of interest (ROI) for ROI extraction and skeletal maturity determination of each ROI
…..…….……..31
Table 14. Test data distribution and skeletal maturity indicator (SMI) prediction accuracy of the system trained for each region of interest (ROI)
……..…….……..32

List of figures

Figure 1. Thirteen hand and wrist bones observed according to TW3 method
……..……..……6
Figure 2. Fishman's skeletal maturity indicators
……..……..……7
Figure 3. Sample distribution according to age and gender
……..……..……..13
Figure 4. Six regions of interest (ROIs) labeled with rectangles using YOLO-mark tool for the Fishman method
..……..……..……14
Figure 5. Overview of data preprocessing
..……..……15
Figure 6. Proposed fully-automated deep learning flow
……..……16-17
Figure 7. Comparison of the matched mean chronological ages with reference to skeletal maturity indicators (SMIs) between this study and Fishman's study
……..……..23
Figure 8. Examples of attention maps for the system trained with the dataset labeled with only skeletal maturity indicators (SMIs)
……..……...37

I. Introduction

Chronological age (CA) has not been regarded as a reliable parameter for growth evaluation due to individual variation of maturational patterns.[1] Although previous studies have reported different developmental indices such as dental, somatic, and sexual maturity for growth assessement,[2-4] methods based on bone age (BA) or skeletal maturity estimated from hand-wrist radiographs are most widely used in terms of many ossification centers, simplicity, and low radiation exposure to patients.[5]

Greulich-Pyle (GP) and Tanner-Whitehouse 3 (TW3) methods have been typically employed for bone age assessment (BAA), but Fishman method is also commonly known as a more concise approach. The GP method[6] consists of a simple comparison with reference images labeled by BAs in a hand atlas. A patient's radiograph is matched with the closest image in a template series, and the labeled BA of the image is then assessed as the BA of the patient. This could be quickly implemented, however, it is difficult to achieve high inter- and intra-observer reliability through reproducible assessment in that all the hand bones in an image would be skimmed through.

In contrast, the TW3 method[7] and the Fishman method[8,9] evaluate the specific bones of the hand and wrist, that is, regions of interest (ROIs). The TW3 method observes thirteen ROIs based on the RUS (radius, ulna and selected short bones) scoring system (Figure 1). The scores are assigned to each ROI following sequential maturation levels. The scores from all ROIs are aggregated and converted into a BA using a score-BA (in 0.1 years) table by gender. This method is rather complicated and requires more time; however, it is more objective and reproducible.[10] The Fishman method observes six ROIs and matches with one of the eleven skeletal maturity indicators (SMIs) based on whether or not a specified ROI reached the defined maturation stage (Figure 2). This method is differenet from the two aforementioned methods because it presents the SMIs instead of BAs in years. Moreover, Fishman described the average CA standards for the eleven SMIs by gender through longitudinal and cross-sectional researches.[8]

BAA is useful for the diagnosis of pediatric endocrine and orthopedic disorders with abnormal growth deviation, in addition to being considered as an individual's age reference

from a forensic viewpoint. Particularly, in clinical dental practice, it is very important to evaluate pubertal growth spurt and residual facial growth for orthodontic treatment planning.[11] In this regard, it would be advantageous that the Fishman's SMIs can give the approximate percentage of adolescent growth completed corresponding to each of the indicators considering the pubertal growth period (Table 1).[12]

Automation techniques for object detection and image classification are being rapidly advanced using convolutional neural networks (CNNs) and their variants, i.e., a specific type of deep learning technologies. BAA from hand-wrist radiographs is one of the ideal tasks that can benefit from artificial neural networks for automated image analysis. In recent years, the studies using deep learning have reported promising results in BAA from hand-wrist radiographs.[5, 13-17]

BoneXpert[18] is a commercially available software, which was introduced in 2009. The software is based on conventional image processing techniques, and had been a typical automated BAA tool until deep learning techniques such as CNNs emerged. BoneXpert is a three-layered system that has been clinically approved and validated for various ethnicities, but it has a critical drawback: it requires a BA-CA relationship as input data for target ethnicities. Furthermore, it tends to be susceptible to image quality degradation such as noise and positioning[14,19] and has relatively low image processing accuracy.[16]

Several deep learning systems for fully automated BAA have been developed by utilizing pre-trained, existing models such as BoNet, GoogLeNet and VGGNet. Of the systems, commercialized software, e.g., VUNOMed-Bone Age[13] was also included. Having begun with these systems, which are based on the GP method that consists of simple comparison and matching is attributed that to obtain accurate ROI localization with conventional segmentation algorithms was difficult and effective deep learning technique for ROI extraction had not been developed yet. However, the first TW3-based BAA system using deep learning was proposed in 2019.[16] The ROI extraction process of this system consists of two steps. Four bounding ROIs (bROIs), which are large areas including actual ROIs, are extracted, and thirteen actual ROIs are then extracted from the bROIs. It is a kind of hybrid technique in that the first step uses convention image processing and the second step uses deep learning techniques. Most of the above-mentioned systems have shown approximation errors of 0.5 years with reference to the ground truth, indicating the

possibility of achieving accurate and effective BAA.

However, it is questionable whether these methods, which were developed based on past American or European population data, can be applied for the BAA of contemporary children of different ethnicities and backgrounds. If the BAs estimated by these methods are significantly different from the corresponding CAs in a limited group of healthy subjects, it is possible to postulate that a kind of stage concept such as SMIs would be more valid and consistent than the BAs in years. In other words, it might be more appropriate to focus on the qualitative skeletal maturity indices such as SMIs, if only the growth level assessment is significant, rather than forensic medicine for which age estimation is necessary. Nevertheless, few comparison studies among BAA methods including the Fishman method, have been reported for contemporary healthy Korean children and adolescents. Besides, to the best of knowledge, no fully-automated deep learning system for skeletal maturity determination based on the Fishman's SMIs has been developed.

Configuration and training outcome of a neural network model could depend on its reliance on the capabilities of deep learning or human interventions. Moreover, even if a system's predicted values are similar to human estimations, there is still a question on the similarity/dissimilarity of the decision-making mechanisms of the two approaches. In other words, the Fishman method factored in the varying accuracy between a system trained with only an SMI-labeled hand-wrist radiograph dataset and another one trained with a dataset that was not only labeled with SMIs but was additionally labeled relating to ROI extraction and maturation status using each ROI. It is also questionable as to whether the former system (i.e., the system trained with only an SMI-labeled data) would focus on the identical regions as the Fishman's ROIs for skeletal maturity determination.

Accordingly, this study first aimed to compare the Fishman's SMIs with the GP and the TW3 methods for BAA from hand-wrist radiographs and assess whether SMIs are a reliable index in contemporary healthy Korean prepubertal and pubertal populations. Two fully-automated deep learning systems for skeletal maturity determination based on the Fishman's SMIs were developed. One system was trained with only an SMI-labeled dataset, and the other one was trained with both SMI- and ROI-labeled dataset. The second can automate the entire skeletal maturity determination process: the extraction of six ROIs, skeletal maturity determination for each of the six extracted ROIs, and final SMI prediction.

To assess the accuracy of the proposed systems, the SMIs produced by each system were compared with the values from a reference standard established by two experienced oral and maxillofacial radiologists.

Table 1. Approximate chronological ages and percentage of growth completed corresponding to skeletal maturity indicators (SMIs)[12]

SMI	Age (year)	Percentage of adolescent growth completed	Percentage of maxillary growth completed	Percentage of mandibular growth completed
Male				
1	11.01±1.22			
2	11.68±1.06	15.0	16.7	15.9
3	12.12±1.00	21.6	18.5	19.5
4	12.33±1.09	28.9	20.3	26.7
5	12.98±1.12	34.0	28.6	30.8
6	13.75±1.06	52.6	49.7	48.5
7	14.38±1.08	74.3	69.0	66.7
8	15.11±1.03	87.3	83.0	77.7
9	15.50±1.07	92.0	89.6	84.6
10	16.40±1.00	95.3	92.7	91.5
11	17.37±1.26	100.0	100.0	100.0
Female				
1	9.94±0.96			
2	10.58±0.88	12.2	16.7	14.7
3	10.88±0.99	22.5	18.5	25.0
4	11.22±1.11	32.7	20.3	33.1
5	11.64±0.90	39.8	28.6	38.3
6	12.06±0.96	51.7	49.7	47.0
7	12.34±0.90	73.6	69.0	58.0
8	13.10±0.87	86.6	83.0	72.7
9	13.90±0.99	91.9	89.6	84.0
10	14.77±0.96	96.1	92.7	90.0
11	16.07±1.25	100.0	100.0	100.0

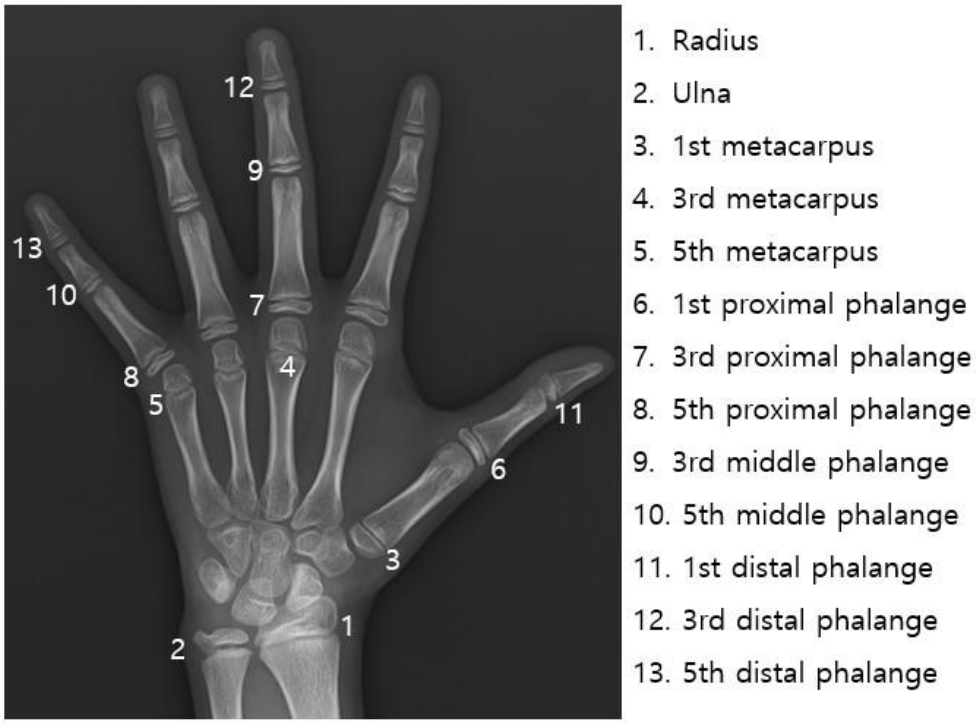
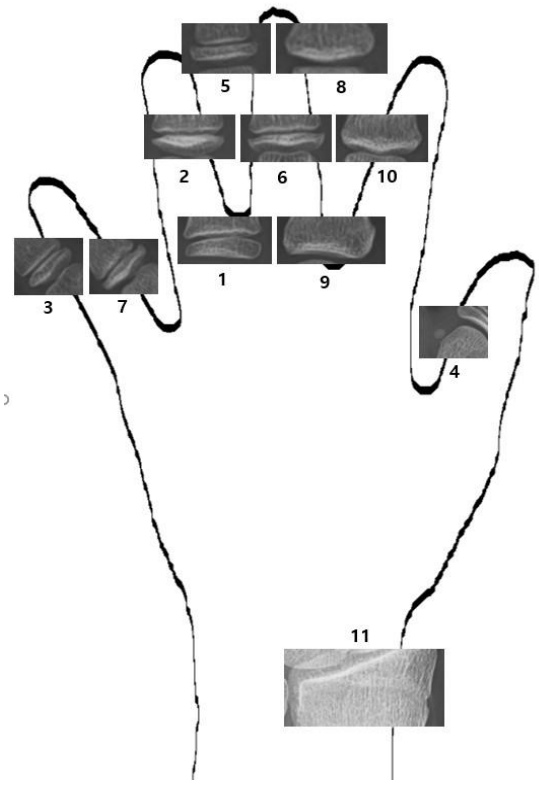


Figure 1. Thirteen hand and wrist bones observed according to TW3 method



I. Width of epiphysis as wide as diaphysis
1. 3rd finger-proximal phalange
2. 3rd finger-middle phalange
3. 5th finger-middle phalange

II. Ossification
4. Adductor sesamoid of thumb

III. Capping of epiphysis
5. 3rd finger-distal phalange
6. 3rd finger-middle phalange
7. 5th finger-middle phalange

IV. Fusion of epiphysis and diaphysis
8. 3rd finger-distal phalange
9. 3rd finger-proximal phalange
10. 3rd finger-middle phalange
11. Radius

Figure 2. Fishman's skeletal maturity indicators[8]

II. Materials and Methods

This study was approved by the Institutional Review Board of Seoul National University Dental Hospital (ERI20011).

1. Data collection

Digital left hand-wrist radiographs were retrospectively selected from the picture archiving and communication system (PACS) of Seoul National University Dental Hospital. All of the radiographs were taken from 2012 to 2017 using REX 650R (Listem Co., Ltd., Wonju, South Korea) under 50 kV and 8 mAs for growth evaluation related to orthodontic treatment. FCR XG5000 (FUJIFILM Co., Tokyo, Japan) was used for image acquisition. The CAs of the subjects were 6–17 years old when the radiographs were taken, and calculated by subtracting the subjects' birth dates from the dates the radiographs were taken. Data were obtained in BMP formats of image file using Infinitt® PACS software (Infinitt Healthcare Co., Ltd., Seoul, South Korea), and the subjects' information was not recorded.

Exclusion criteria were as follow: 1) systemic disease such as developmental or endocrinological disorder, 2) bony abnormality of hand and wrist region due to trauma or disease, and 3) inappropriate radiograph: poor image quality, poor position, or patient movement. A total of 1,617 radiographs (706 males and 911 females) were collected for the study. Figure 3 and Table 2 show the data distribution according to age and gender.

2. Bone age assessment

For the GP method, an open deep learning system by 16 Bit (https://www.16bit.ai/bone-age) was used. The system achieved a mean absolute difference of 4.265 months and concordance correlation coefficient of 0.991, placing the system at 1st position in the RSNA 2017 Machine Learning Challenge.[20]

For the TW3 and the Fishman methods, BAA was performed by two observers: 4- and 7-year-experienced oral and maxillofacial radiologists. For the TW3 method, BA estimation was done for 200 samples (100 samples for male and female, respectively),

which were selected randomly but distributed as evenly as possible by the CAs. For the Fishman method, the first observer estimated SMIs for all of 1,617 images and the second observer did for 600 images (300 images for male and female, respectively), which were selected randomly, and the BAs were determined based on the mean CA value corresponding to each SMI as shown in Table 1.

All radiographs were evaluated on a diagnostic display screen (Nio Color 2MP LED 21.3-inch monitor with 1,200 x 1,600 resolution; BARCO Ltd., Seoul, South Korea). The evaluation was repeatedly performed by the observers with a 3-week interval. The observers were unaware of each other's assessments, as well as of their first estimation during the second assessment. A consensus was reached on the SMI reference standard, which will be used in training the proposed systems, through discussion in the event of a disagreement.

3. Data preparation for deep learning

3.1. Labeling of the image data

First, all of the radiographs were labeled with at least one of Fishman's eleven SMIs or SMI 0. SMI 0 referred to a status with less skeletal maturity than SMI 1. For ROI extraction, all images were annotated with rectangular boxes using YOLO-mark tool, denoting each of six ROIs with numbers (Figure 4). Additionally, to label the skeletal maturity of the ROIs, defined skeletal maturation stages for each ROI were applied according to characteristics associated with the Fishman's SMI evaluation (Table 3) and combination of the ROI- based stages for each SMI (Table 4).

3.2. Data distribution and augmentation

The sample radiographs in the dataset were randomly divided at a ratio of 7:2:1 for training, validation, and testing. The sizes of the training and validation set were increased through data augmentation to reduce overfitting and acquire high accuracy. The radiographs were rotated from -9° to +9° with an 1° interval, and their intensities were transformed to 0.7 and 1.3 times, respectively. Hence, 56 images were generated from each radiograph, totaling more than 66,000 images.

3.3. Data preprocessing

3.3.1. Masking

Since the radiographic images contains irrelevant values such as noise, which interfere with deep learning, the hand and wrist areas of the images were masked (Figure 5). Hand-wrist radiographs were converted to texture images using a range filter, and masked hand-wrist images were extracted by opening method among morphological structure techniques[21] using MATLAB® (MathWorks, Natick, MA, USA). By combining the extracted masked images and the original images, the pixel values of the remaining areas except for hand and wrist areas were adjusted to zero so as not to affect the learning process.

3.3.2. Histogram equalization

It is necessary to achieve a radiopacity that is appropriate for the systems to display the data images because the original radiographic images may have higher brightness with lower contrast values. Therefore, a histogram equalization technique was applied to expand the high values (white pixels) and compress the values in the dark layer (Figure 5). This allows the images to be displayed with more details by stretching the histogram. This can eliminate low contrast in images and improve image quality.[22] However, since applying the technique to the whole image data, which consist of light and dark parts, is not useful, contrast-limited adaptive histogram equalization was applied. In other words, each image was divided into smaller tiles (8×8), and the method was applied to each tile. Contrast limits were set to avoid noise amplification because noise (extreme dark or light parts) in any small area can adversely affect the transformed image.

4. Proposed systems and training details

The deep learning systems were implemented on four graphics processing units (GeForce GTX 1080 Ti, 11 GB VRAM, NVIDIA Co., Santa Clara, CA, USA). At first, a modified neural network based on ResNet50[23] was developed for classifying and identifying SMIs in a hand-wrist radiograph. ResNet50 is a deep neural network equipped with state-of-the-art residual blocks and improved feature representation ability. The training data had a

resolution of 2,000 x 2,510 pixels. The network was trained on 100 epochs with a 4-batch size, and comprises stochastic gradient descent for optimization, categorical cross entropy for loss function, $1e^{-6}$ decay, and momentum of 0.9.

A modified version of YOLOv3[24], which is a single shot object detection method, was applied to extract the six ROIs from the input image data. Six different classification models based on ResNet50 were developed to evaluate the skeletal maturation of each ROI. Each model was used to classify and identify the unique feature that indicates the skeletal maturity level for each ROI. The final SMIs were obtained by aggregating the predictions of the six models.

Detailed configuration and flow of the proposed fully-automated deep learning systems are shown in Figure 6.

5. Accuracy evaluation of the trained systems using the test dataset

To evaluate ROI extraction accuracy, recall (sensitivity), precision (positive predictive value), and F1-score were calculated, as shown below.

$$\text{recall} = \frac{\text{TP}}{\text{FN+TP}}$$

$$\text{precision} = \frac{\text{TP}}{\text{FP+TP}}$$

$$\text{F1-score} = \frac{2\times\text{precision}\times\text{recall}}{\text{precision+recall}}$$

where TP represents true positive, FN represents false negative, and FP represents false positive. The average precision value was measured as the average value of precision across all recall values. These are common parameters in object detection.[25,26]

Skeletal maturity prediction accuracy for each ROI was evaluated with the concordance rate (%) between the predicted stages from the systems and the corresponding stages from the reference standard.

The SMI prediction accuracy of the trained deep learning systems was measured using mean absolute error (MAE) and concordance rate (%) based on the reference standard. The formula for the MAE is given below.

$$\text{MAE} = \frac{1}{n}\sum_{j=1}^{n}\left|y_j - \hat{y}_j\right|$$

where n represents the number of test images, and $\hat{y}j$ and yj denote the predicted SMI by the systems and the corresponding ground truth from the reference standard, respectively. The concordance rate was subdivided into top-1, within-1, and within-2 prediction accuracies. The top-1, within-1, and within-2 denote the concordance rate between the predicted SMI by the systems and the ground truth, the rate within one of the ground truth, and the rate within two of the ground truth, respectively, where the ground truth is the corresponding SMI from the reference standard.

6. Statistical analysis

The samples in this study satisfied the normal distribution for all BAA methods. Paired t-test ($p < 0.05$) and Pearson (for the GP and TW3 methods) or Spearman (for the Fishman method) correlation analysis ($p < 0.05$) were used to compare the BAs estimated by each BAA method with the CAs. To assess the intra- and inter-observer reliability for the TW3 and Fishman methods, Cohen's Kappa coefficients were calculated. IBM SPSS statistics 23 (SPSS Inc., Chicago, IL, USA) was used for statistical calculations. The values related to the accuracy of the trained skeletal maturity determination systems were calculated using Excel 2019 (Microsoft, Redmond, Wash, USA).

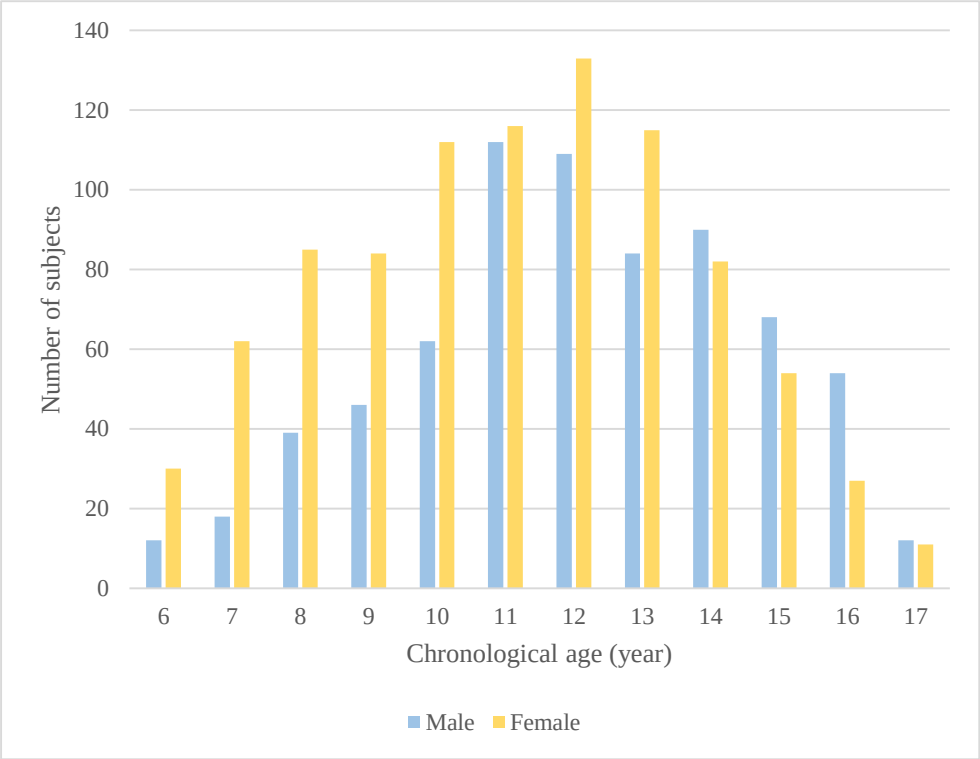


Figure 3. Sample distribution according to age and gender

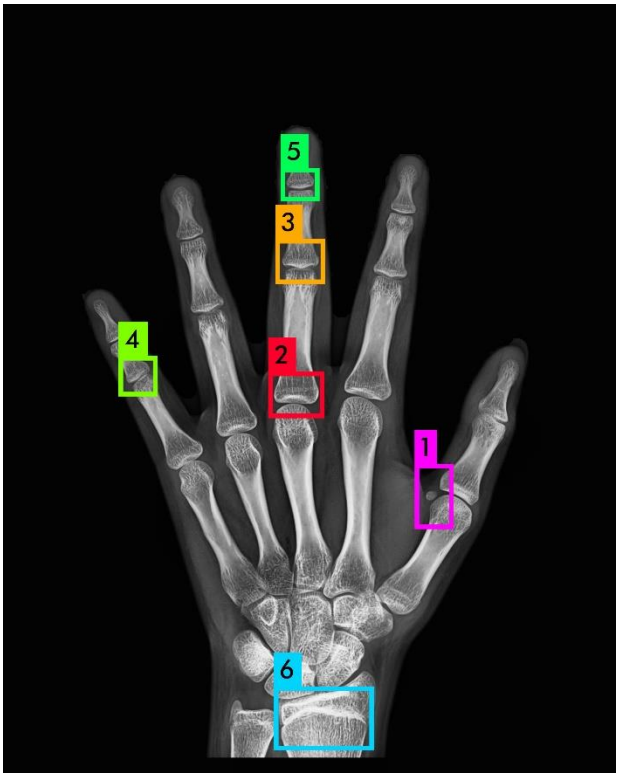


Figure 4. Six regions of interest (ROIs) labeled with rectangles using YOLO-mark tool for the Fishman method. ROI 1: adductor sesamoid of thumb; ROI 2: proximal phalange of 3rd finger; ROI 3: middle phalange of 3rd finger; ROI 4: middle phalange of 5th finger; ROI 5: distal phalange of 3rd finger; ROI 6: radius.

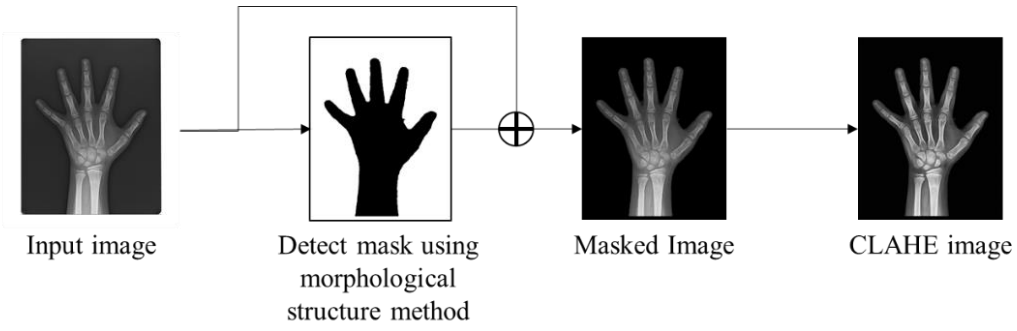
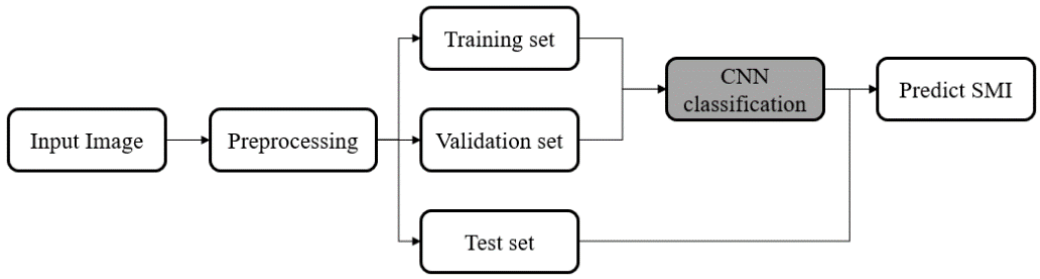


Figure 5. Overview of data preprocessing. Masking of the hand and wrist area was performed to remove irrelevant values from input data. By combining the extracted masked image and the original image, pixel values of the remaining areas except for the hand and wrist area changed to zero. Additionally, a contrast-limited adaptive histogram equalization (CLAHE) technique was used to obtain final images, which has the radiographic contrast appropriate for deep learning.

a)



b)

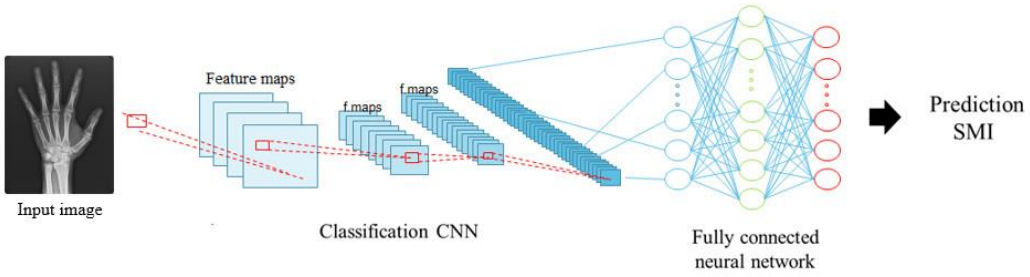
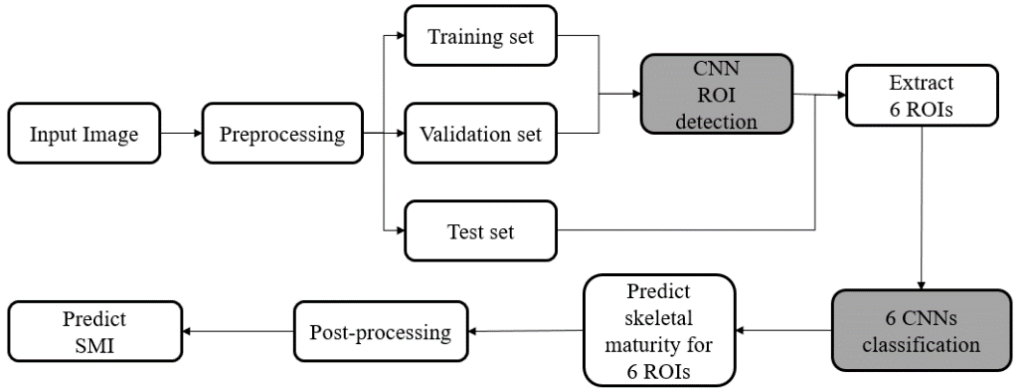
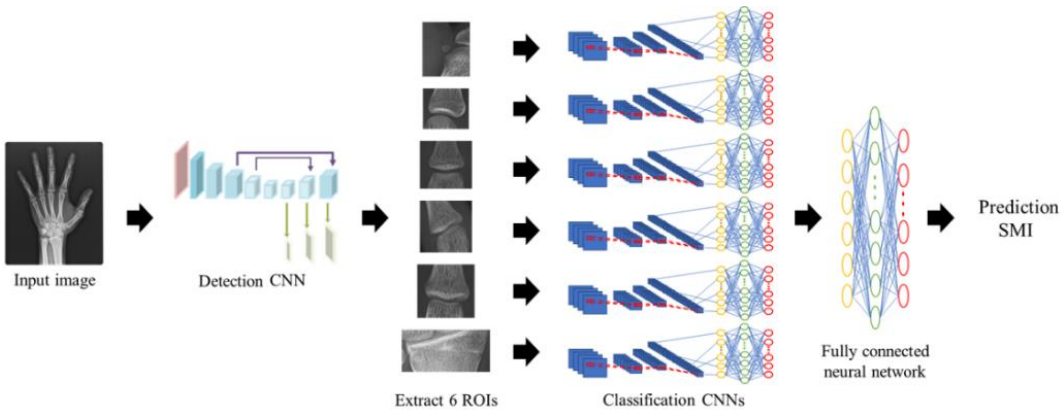


Figure 6. Proposed fully-automated deep learning flow. a) and b) System trained with the data labeled with only skeletal maturity indicators (SMIs). c) and d) System trained with data labeled for regions of interest (ROI) extraction, skeletal maturity prediction for each ROI, and final SMI prediction. e) Configuration of a ResNet50-based neural network for the final SMI prediction.

c)



d)



e)

	Layer 1	Layer 2	Layer 3	Layer 4	Layer 5	Layer 6	Fully connected neural network
Our model	[7x7,64 stride 2 3x3, max pool stride 2]	[1x1, 64 3x3, 64 1x1, 1256] x3	[1x1, 1128 3x3, 3128 1x1, 1512] x4	[1x1, 1256 3x3, 3256 1x1,11024] x5	[1x1, 1512 3x3, 3512 1x1,12048] x6	[Average pool 2048-d FC ReLU 2,3 or 4-d FC SoftMax]	[64-d FC ReLU 32-d FC ReLU 12-d FC SoftMax]

Figure 6. Proposed fully-automated deep learning flow. a) and b) System trained with the data labeled with only skeletal maturity indicators (SMIs). c) and d) System trained with data labeled for regions of interest (ROI) extraction, skeletal maturity prediction for each ROI, and final SMI prediction. e) Configuration of a ResNet50-based neural network for the final SMI prediction.

Table 2. Sample distribution according to age and gender

Age (years)	Male	Female	Total
6	12	30	42
7	18	62	80
8	39	85	124
9	46	84	130
10	62	112	174
11	112	116	228
12	109	133	242
13	84	115	199
14	90	82	172
15	68	54	122
16	54	27	81
17	12	11	23
Total	706	911	1617

Table 3. Skeletal maturation stages for each region of interest (ROI) defined in the study

ROI	Characteristics		Skeletal maturation stage
1	Ossification of the adductor sesamoid	No	0
		Yes	1
2	Width of epiphysis and diaphysis	Not equal	0
		Equal	1
	Fusion of epiphysis and diaphysis	No	
		Yes	2
3	Width of epiphysis and diaphysis	Not equal	0
		Equal	1
	Capping of epiphysis	No	
		Yes	2
	Fusion of epiphysis and diaphysis	No	
		Yes	3
4	Width of epiphysis and diaphysis	Not equal	0
		Equal	1
	Capping of epiphysis	No	
		Yes	2
5	Capping of epiphysis	No	0
		Yes	1
	Fusion of epiphysis and diaphysis	No	
		Yes	2
6	Fusion of epiphysis and diaphysis	No	0
		Yes	1

Table 4. Combination of skeletal maturation stages of regions of interests (ROIs) for each skeletal maturity indicator (SMI)

SMI	ROI 1	ROI 2	ROI 3	ROI 4	ROI 5	ROI 6
0	0	0	0	0	0	0
1	0	1	0	0	0	0
2	0	1	1	0	0	0
3	0	1	1	1	0	0
4	1	1	1	1	0	0
5	1	1	1	1	1	0
6	1	1	2	1	1	0
7	1	1	2	2	1	0
8	1	2	2	2	1	0
9	1	2	2	2	2	0
10	1	2	3	2	2	0
11	1	2	3	2	2	1

III. Results

The BAs measured with the GP, TW3 and Fishman methods were analyzed based on CAs and gender. Tables 5–7 present the mean CAs and BAs, and the differences between them for each method. For the Fishman method, 117 males (6-12 years) and 87 females (6-10 years) were excluded from this evaluation since they showed less skeletal maturity than SMI 1. This stage was not assigned for the corresponding average standard CA by Fishman, and defined as SMI 0 in the study as above mentioned. There were statistically significant differences ($p < 0.05$) between the CAs and the BAs in both the overall group and gender subgroups for all the methods except in the male group for the TW3 method ($p = 0.839$). In addition, in general, the BAs showed an overestimation trend compared to the CAs for all three methods, suggesting faster growth in the subjects in this study than in the populations on which these methods were based.

However, significant correlations were observed between the CAs and BAs in both genders for all three methods as shown in Table 8 ($p = 0.000$). The Spearman correlation coefficients of the Fishman's SMIs with CAs showed approximately equivalent values with the Pearson correlation coefficients of the GP- or TW3-based BAs with CAs for both genders, suggesting that SMIs are a reliable index for contemporary Korean populations. The analysis for the SMIs was not performed in the overall group since the age distribution by SMIs varies depending on the gender.

For the inter-and intra-observer reliability, the Fishman method was superior to the TW3 method (Table 9). The Fishman's SMIs resulted in almost perfect level of kappa coefficients for both inter- and intra-observer reliability, indicating that this SMI labeling is sufficient and reliable to be used as a reference standard for training deep learning system for skeletal maturity determination.

The sample distribution based on SMIs and the mean CAs corresponding to each SMI in this study compared to Fishman's study was shown in Figure 7 and Table 10. The mean CA difference for each SMI between this study and Fishman's study averaged 0.72 and 0.91 years in the male and female, respectively, showing the lowering trend on the whole SMIs for both genders. The mean CA interval between the SMIs in this study was not quite

different from that in Fishman's study. In this study, a slightly lower value of 0.58 years in males was obtained compared to that Fishman's study, which is 0.64 years. On the other hand, 0.68 years in females was obtained in this study, which was higher than the Fishman's 0.61 years (Table 11).

The mean CA of all the subjects used for developing the BAA systems was 12.01±2.62 (range, 6.04-17.99 years). 204 (male: 117, female: 87) radiographs that showed less skeletal maturity than SMI 1 were labeled as SMI 0 and included in the dataset to train the deep learning systems.

The MAE in SMI prediction for the system trained with only SMIs was 0.46, 1.24 and 0.88 in males, females, and overall, respectively. The concordance rate of the system was shown in Table 12. The top-1 prediction accuracy was 61.7 %, 28.1 %, and 43.5 %, and the within-1 prediction accuracy was 96.3 %, 53.3 %, and 73.1 % in males, females, and overall, respectively, indicating that the values in males were superior to those in females. Moreover, SMI in the female group was widely varied.

Meanwhile, the outcome of the system consisting of ROI extraction, skeletal maturation determination for each ROI and final SMI prediction was described in Table 13 and 14. The ROI extraction accuracy was above 90.0 % for all six ROIs, and approximately 100.0 % for ROI 1. The skeletal maturity determination accuracy for each ROI was also high, which was 88.3 % on average, particularly 100.0 % for ROI 1. The final SMI prediction accuracy outperformed that of the system trained with only SMIs. The MAE was 0.39, 0.30 and 0.34, the top-1 prediction accuracy was 75.0 %, 79.6 % and 77.6 %, and the within-1 prediction accuracy was 90.8 %, 95.9 % and 93.7 % in males, females, and overall, respectively. The system showed slightly better values in females than in males, and less deviation of the accuracies for SMI compared to the system without ROI-based training.

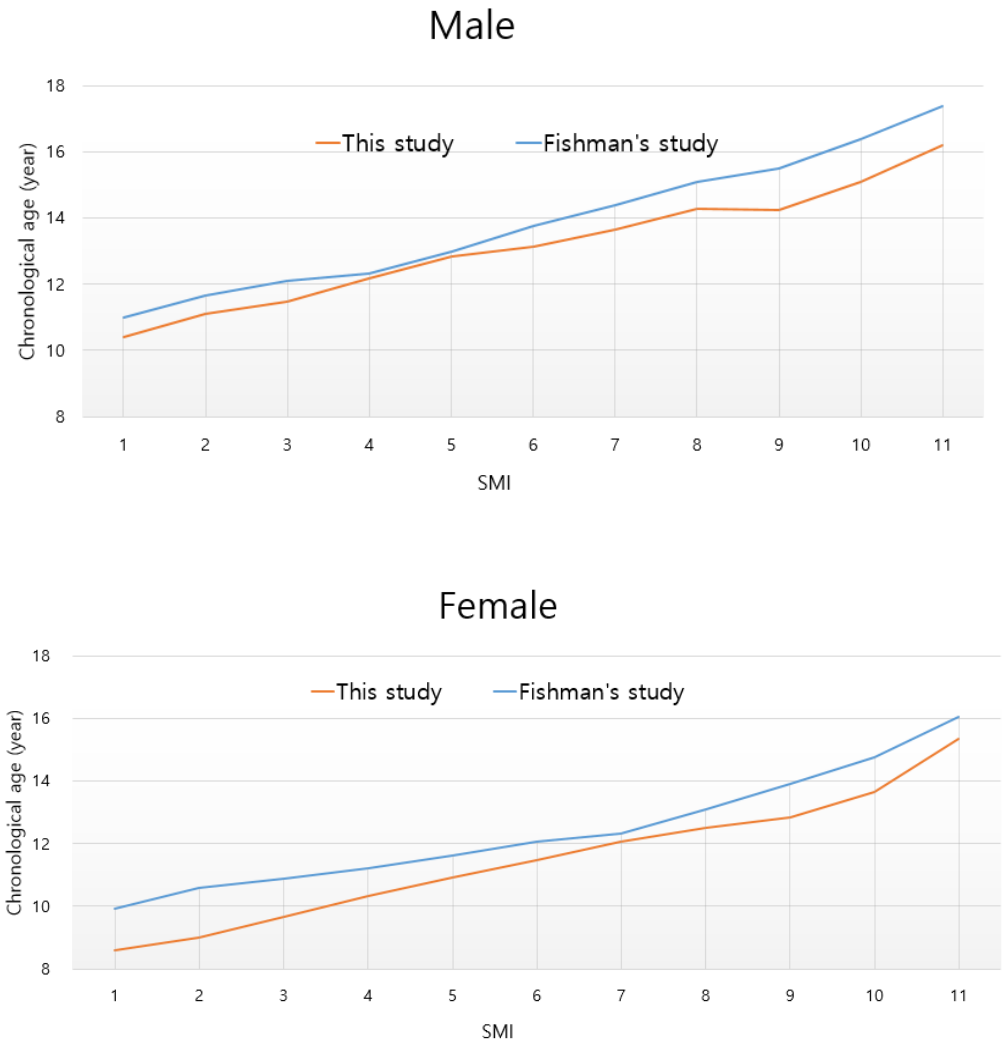


Figure 7. Comparison of the matched mean chronological ages with reference to skeletal maturity indicators (SMIs) between this study and Fishman's study[8]

Table 5. Means and standard deviations of chronological ages (CAs) and bone ages (BAs) estimated by GP method in each age group (unit: year)

Age	Male				Female				Overall			
	No.	Mean CA	Mean BA	Mean dif.	No.	Mean CA	Mean BA	Mean dif.	No.	Mean CA	Mean BA	Mean dif.
6	12	6.60±0.27	6.58±0.96	-0.02±0.83	30	6.56±0.30	6.73±0.97	0.17±0.92	42	6.57±0.29	6.69±0.97	0.12±0.89
7	18	7.62±0.30	8.11±1.10	0.49±1.10	62	7.45±0.27	7.56±0.93	0.11±0.93	80	7.49±0.29	7.69±1.01	0.19±0.95
8	39	8.49±0.31	8.71±1.32	0.22±1.22	85	8.51±0.31	8.48±1.03	-0.03±0.93	124	8.50±0.31	8.55±1.13	0.05±1.02
9	46	9.52±0.27	10.27±1.43	0.75±1.33*	84	9.46±0.27	9.46±1.05	0.00±0.99	130	9.48±0.27	9.75±1.25	0.26±1.18*
10	62	10.56±0.27	11.79±1.31	1.23±1.25*	112	10.53±0.29	10.72±0.96	0.19±0.91*	174	10.54±0.28	11.10±1.22	0.56±1.16*
11	112	11.47±0.27	12.36±1.08	0.89±1.01*	116	11.48±0.27	12.09±1.23	0.61±1.22*	228	11.48±0.27	12.22±1.17	0.75±1.12*
12	109	12.54±0.30	13.46±0.87	0.92±0.82*	133	12.50±0.28	13.40±1.36	0.90±1.29*	242	12.52±0.29	13.43±1.17	0.91±1.10*
13	84	13.44±0.31	14.14±1.26	0.70±1.25*	115	13.46±0.29	14.36±1.06	0.80±1.23*	199	13.46±0.30	14.27±1.15	0.81±1.13*
14	90	14.53±0.30	15.32±1.26	0.79±1.23*	82	14.50±0.28	15.04±0.50	0.54±0.53*	172	14.52±0.29	15.19±0.99	0.68±0.96*
15	68	15.49±0.32	16.14±1.06	0.65±0.98*	54	15.46±0.27	15.34±0.48	-0.12±0.54	122	15.48±0.30	15.79±0.94	0.31±0.90*
16	54	16.48±0.27	17.02±0.37	0.54±0.40*	27	16.43±0.30	15.55±0.70	-0.88±0.81*	81	16.47±0.28	16.53±0.86	0.07±0.88
17	12	17.46±0.30	17.12±0.33	-0.34±0.26*	11	17.37±0.30	15.76±0.45	-1.61±0.49*	23	17.42±0.31	16.47±0.80	-0.95±0.75*
Total	706	12.58±2.52	13.33±2.74	0.75±1.09*	911	11.59±2.61	11.94±2.88	0.35±1.12*	1617	12.02±2.62	12.55±2.90	0.52±1.11*

dif.: Difference, *: $p < 0.05$

Table 6. Means and standard deviations of chronological ages (CAs) and bone ages (BAs) estimated by TW3 method in each age group (unit: year)

Age	Male				Female				Overall			
	No.	Mean CA	Mean BA	Mean dif.	No.	Mean CA	Mean BA	Mean dif.	No..	Mean CA	Mean BA	Mean dif.
7	9	7.58±0.33	6.78±0.71	-0.80±0.86*	11	7.29±0.21	7.85±1.70	0.56±1.67	20	7.42±0.30	7.37±1.45	-0.03±1.49
8	11	8.41±0.25	7.41±1.32	-1.00±1.28*	11	8.51±0.34	8.75±1.71	0.26±1.49	22	8.46±0.64	8.60±0.26	-0.41±1.51
9	11	9.45±0.23	8.65±1.06	-0.80±1.04*	11	9.51±0.33	10.27±0.85	0.80±0.82*	22	9.48±0.28	9.48±1.24	-0.02±1.23
10	11	10.38±0.33	10.15±1.63	0.23±1.30	11	10.46±0.30	11.26±0.70	0.86±0.20*	22	10.42±0.31	10.74±1.21	0.30±1.14
11	11	11.46±0.28	12.19±1.02	0.73±0.86*	11	11.55±0.27	12.72±0.80	1.20±0.77*	22	11.51±0.58	12.51±0.19	0.97±0.82*
12	12	12.53±0.29	13.05±1.06	0.52±1.09	12	12.57±0.33	13.00±1.54	0.48±1.36	24	12.55±0.30	13.12±1.22	0.54±1.22*
13	11	13.47±0.29	13.56±1.39	0.09±1.37	11	13.31±0.24	14.02±0.69	0.76±0.64*	22	13.39±0.27	13.85±1.06	0.45±1.08
14	11	14.61±0.19	15.12±0.85	0.51±1.01	11	14.37±0.27	14.47±0.23	0.14±0.16*	22	14.49±0.26	14.88±0.75	0.37±0.75*
15	13	15.53±0.32	16.08±0.29	0.55±0.31*	11	15.42±0.31	14.77±0.26	-0.65±0.42*	24	15.48±0.31	15.50±0.72	0.01±0.71
Total	100	11.66±2.61	11.69±3.36	0.03±1.23	100	11.45±2.58	11.90±2.59	0.50±1.10*	200	11.56±2.60	11.81±2.96	0.24±1.18*

dif.: Difference, *: $p < 0.05$

Table 7. Means and standard deviations of chronological ages (CAs) and bone ages (BAs) estimated by Fishman method in each age group (unit: year)

Age	Male				Female				Overall			
	No.	Mean CA	Mean BA	Mean dif.	No.	Mean CA	Mean BA	Mean dif.	No.	Mean CA	Mean BA	Mean dif.
6	-	-	-	-	5	6.81±0.18	10.20±0.35	3.43±0.31*	5	6.81±0.18	10.20±0.35	3.43±0.31*
7	-	-	-	-	29	7.48±0.28	10.31±0.39	2.82±0.44*	29	7.48±0.28	10.31±0.39	2.82±0.44*
8	3	8.84±0.16	11.23±0.39	2.62±0.28*	63	8.57±0.29	10.49±0.49	2.00±0.48*	66	8.59±0.29	10.52±0.51	1.94±0.47*
9	26	9.59±0.25	11.61±0.48	2.04± 0.52*	77	9.48±0.27	10.80±0.53	1.46±0.56*	103	9.51±0.27	11.04±0.61	1.53±0.59*
10	49	10.58±0.27	11.93±0.52	1.42± 0.64*	111	10.54±0.28	11.36±0.58	1.09±0.62*	160	10.55±0.28	11.53±0.62	0.98±0.60*
11	97	11.50±0.26	12.22±0.62	0.81± 0.72*	117	11.48±0.28	12.27±1.02	0.85±0.77*	214	11.49±0.27	12.24±0.86	0.76±0.85*
12	107	12.54±0.30	13.05±1.04	0.82±0.94*	133	12.50±0.28	13.35±1.27	0.79±1.07*	240	12.52±0.29	13.22±1.18	0.70±1.11*
13	83	13.45±0.31	13.86±1.50	0.63± 1.28*	115	13.46±0.29	14.34±1.10	0.67±0.93*	198	13.46±0.30	14.14±1.30	0.68±1.28*
14	90	14.53±0.30	15.22±1.51	0.64± 1.14*	82	14.50±0.28	15.14±0.84	0.30±0.73*	172	14.52±0.29	15.18±1.23	0.66±1.21*
15	68	15.49±0.32	16.11±1.12	0.44± 0.99*	54	15.46±0.27	15.61±0.63	0.13±0.71*	122	15.48±0.30	15.89±0.96	0.41±0.92*
16	54	16.48±0.27	17.08±0.51	0.61± 0.51*	27	16.43±0.31	15.88±0.75	-0.74±0.91*	81	16.47±0.28	16.68±0.82	0.21±0.84*
17	12	17.46±0.31	17.21±0.38	-0.49± 0.38*	11	17.37±0.31	16.07±0.00	-1.54±0.49*	23	17.42±0.31	16.66±0.64	-0.75±0.61*
Total	589	13.29±2.03	14.00±2.07	-0.81(1.04)*	824	12.02±2.36	12.93±2.00	0.96±1.13*	1413	12.55±2.31	13.38±2.10	0.83±1.10*

dif.: Difference, *: $p < 0.05$

Table 8. Correlation coefficients between chronological age and bone age (GP method and TW3 method) or skeletal maturity indicator (Fishman method) measured by each method

Group	GP method	TW3 method	Fishman method
Male	0.917*	0.946*	0.900**
Female	0.925*	0.910*	0.927**
Overall	0.924*	0.924*	-

*: p = 0.000 (Pearson correlation test)

**: p = 0.000 (Spearman correlation test)

Table 9. Inter- and intra-observer reliability for TW3 method and Fishman method

k	Inter-observer	Intra-observer 1	Intra-observer 2
TW3 method	0.750	0.846	0.817
Fishman method	0.843	0.912	0.819

k: Cohen's kappa coefficient

Table 10. Sample distribution with reference to skeletal maturity indicator (SMI) and comparison between the matched mean chronological ages (CAs) in this study and that in Fishman's study (unit: year)

	Male				Female			
SMI	No.	Mean CA in this study	Mean CA in Fishman's study	Dif.	No.	Mean CA in this study	Mean CA in Fishman's study	Dif.
1	26	10.42±1.16	11.01±1.22	0.59	61	8.61±1.17	9.94±0.96	1.33
2	58	11.09±1.02	11.68±1.06	0.59	62	9.00±1.19	10.58±0.88	1.58
3	60	11.49±1.15	12.12±1.00	0.63	53	9.67±1.10	10.88±0.99	1.21
4	96	12.19±1.12	12.33±1.09	0.14	75	10.34±0.95	11.22±1.11	0.88
5	46	12.83±0.96	12.98±1.12	0.15	74	10.93±0.97	11.64±0.90	0.71
6	37	13.14±1.12	13.75±1.06	0.61	30	11.49±0.72	12.06±0.96	0.57
7	60	13.65±0.98	14.38±1.08	0.73	89	12.08±0.99	12.34±0.90	0.26
8	29	14.30±1.18	15.11±1.03	0.81	42	12.50±0.91	13.10±0.87	0.60
9	14	14.25±0.95	15.50±1.07	1.25	46	12.84±0.92	13.90±0.99	1.06
10	87	15.10±0.98	16.40±1.00	1.30	175	13.66±1.03	14.77±0.96	1.11
11	76	16.22±0.84	17.37±1.26	1.15	117	15.37±1.16	16.07±1.25	0.70
Total	589		Mean dif.	0.72	824		Mean dif.	0.91

Dif.: Difference

Table 11. Comparison between the mean age intervals based on skeletal maturity indicator (SMI) in this study and that in Fishman's study (unit: year)

SMI	Male		Female	
	This study	Fishman's study	This study	Fishman's study
1-2	0.67	0.67	0.39	0.64
2-3	0.40	0.44	0.67	0.30
3-4	0.70	0.21	0.67	0.34
4-5	0.64	0.65	0.59	0.42
5-6	0.31	0.77	0.56	0.42
6-7	0.51	0.63	0.59	0.28
7-8	0.65	0.73	0.42	0.76
8-9	-0.05	0.39	0.34	0.80
9-10	0.85	0.90	0.82	0.87
10-11	1.12	0.97	1.71	1.30
Mean	0.58	0.64	0.68	0.61

Table 12. Test data distribution and skeletal maturity indicator (SMI) prediction accuracy of the system trained with only SMIs

SMI	Male				Female				Overall			
	No.	Top-1(%)	Within-1(%)	Within-2(%)	No.	Top-1(%)	Within-1(%)	Within-2(%)	No.	Top-1(%)	Within-1(%)	Within-2(%)
0	7	100.0	100.0	100.0	3	66.7	100.0	100.0	10	90.0	100.0	100.0
1	5	100.0	100.0	100.0	7	57.1	57.1	57.1	12	75.0	75.0	75.0
2	6	100.0	100.0	100.0	6	83.3	100.0	100.0	12	91.7	100.0	100.0
3	10	50.0	100.0	100.0	8	0.0	100.0	100.0	18	27.8	100.0	100.0
4	5	100.0	100.0	100.0	4	0.0	25.0	75.0	9	55.6	66.7	88.9
5	5	80.0	100.0	100.0	3	0.0	0.0	33.3	8	50.0	62.5	75.0
6	9	33.3	100.0	100.0	9	11.1	11.1	33.3	18	22.2	55.6	66.7
7	6	66.7	100.0	100.0	10	20.0	40.0	60.0	16	37.5	62.5	75.0
8	2	50.0	100.0	100.0	5	20.0	20.0	60.0	7	28.6	42.9	71.4
9	3	0.0	100.0	100.0	7	42.9	71.4	85.7	10	30.0	80.0	90.0
10	11	0.0	55.6	100.0	17	23.5	52.9	82.4	28	14.3	54.0	89.3
11	5	60.0	100.0	100.0	8	12.5	62.5	87.5	13	30.8	76.9	92.3
Total	74	61.67	96.3	100.0	87	28.1	53.3	72.9	161	43.5	73.1	85.3

Table 13. Accuracy of the system trained using each region of interest (ROI) for ROI extraction and skeletal maturity determination of each ROI

ROI	Accuracy of ROI extraction (%)	Accuracy of skeletal maturity determination for each ROI (%)
1	99.4	100.0
2	98.8	87.4
3	98.5	84.5
4	97.4	70.7
5	94.2	91.4
6	93.7	96.0
Mean	97.0	88.3

Table 14. Test data distribution and skeletal maturity indicator (SMI) prediction accuracy of the system trained for each region of interest (ROI)

SMI	Male				Female				Overall			
	No.	Top-1(%)	Within-1(%)	Within-2 (%)	No.	Top-1(%)	Within-1(%)	Within-2(%)	No.	Top-1(%)	Within-1(%)	Within-2(%)
0	4	100.0	100.0	100.0	17	100.0	100.0	100.0	21	100.0	100.0	100.0
1	6	16.7	83.3	100.0	4	0.0	100.0	100.0	10	10.0	90.0	100.0
2	5	80.0	80.0	100.0	8	87.5	100.0	100.0	13	84.6	92.3	100.0
3	6	50.0	100.0	100.0	7	28.6	100.0	100.0	13	38.5	100.0	100.0
4	7	85.7	85.7	85.7	11	90.9	90.9	90.9	18	88.9	88.9	88.9
5	8	62.5	87.5	100.0	5	100.0	100.0	100.0	13	76.9	92.3	100.0
6	6	83.3	100.0	100.0	2	50.0	100.0	100.0	8	75.0	100.0	100.0
7	11	63.6	72.7	72.7	5	60.0	60.0	80.0	16	62.5	68.8	75.0
8	4	100.0	100.0	100.0	4	100.0	100.0	100.0	8	100.0	100.0	100.0
9	2	100.0	100.0	100.0	5	80.0	80.0	100.0	7	85.7	85.7	100.0
10	10	90.0	100.0	100.0	17	94.1	100.0	100.0	27	92.6	100.0	100.0
11	7	100.0	100.	100.0	13	69.2	100.0	100.0	20	80.0	100.0	100.0
Total	76	75.0	90.8	94.7	98	79.6	96.0	98.0	174	77.6	93.7	96.6

IV. Discussion

This study resulted that the estimated BAs using the GP, TW3 and Fishman methods mostly showed significant differences from the CAs in contemporary healthy Korean children and adolescents. This is consistent with various prior researches,[1, 27] but there have been the studies that found no significant difference between estimated BAs and CAs.[28] In addition, while many studies in other countries showed an overestimation trend in BAs compared to CAs, which generally corresponds with this study,[1] several studies found an underestimation trend.[29,30] This showed the ethnicity, nation or generation variations in skeletal maturation,[31] indicating it may be undesirable to use these methods, which were developed based on past European or American population data, for the BAA of contemporary Korean subjects. In a situation where there is no well-established method for people from all geographic locations, the Fishman method could be used because it has a relatively good consistency for growth evaluation in that it provides the stage of skeletal maturation, not BAs in years.

Fishman described the percentages of adolescent, maxillary and mandibular growth completed corresponding to each of the eleven SMIs.[12] In this regard, SMI 4 is attained at the onset of pubertal growth spurt, SMI 5-7 during the peak velocity, and SMI 11 at the growth completion. This is relatively consistent with the studies of Hägg et al.[32] and Bjork et al.[33]. While maxillary growth showed an approximate identical trend with adolescent growth, mandible resulted in a relatively regular velocity in growth. Additionally, Al-Jewair et al.[34] recently presented a longitudinal study of the American population that both height and mandibular growth were more correlated with BAs than CAs, and BA or height growth was more associated with the growth of mandible than that of maxilla. These were useful for predicting residual height and maxillary/mandibular growth potential as well as for determining the timing of jaw growth control and selecting dental orthodontic treatment or post-growth orthognathic surgery.

However, it is difficult to exclude the possibility of varied individual growth patterns. This may be one of the reasons why there could be a significant discrepancy among treatment outcomes obtained from the same orthodontic device under a similar diagnosis

in clinical practice. In this regard, longitudinal research studies are required to analyze the growth pattern for a large sample in various countries and ethnicities.

One of the important things to consider when comparing BAA methods is inter- and intra-observer variation, which is well-recognized.[35] This variation should be essential when the BAs interpreted by radiologists are used as a reference standard for ground truth to develop automated BAA systems. In this study, almost perfect inter- and intra-observer reliability level was obtained for the Fishman method, which was regarded to be significantly valid for training deep learning systems for skeletal maturity determination. In this study, it may be assumed that the observers' years of work experience as radiologists are rather not sufficient, however, they participated in skeletal maturity evaluation for the study after sufficient interpretation practice for the TW3 and Fishman methods.

Initially, several neural networks were evaluated for SMI prediction. They included VGG16 and ResNet50, which were known as optimal networks for feature extraction and image classification. Although VGG16 achieved higher accuracies for some SMIs after training with SMI-labeled dataset, ResNet50 was finally chosen because of the relatively high communication/computation ratio, which could allow ResNet-based networks to be widely deployed in high-performance applications. Meanwhile, since the current network seems to have excessive parameters regarding the number of the classes and the size of input images, it may be possible to increase the accuracy and efficiency of a deep learning system using a light-weight network.

This study achieved a system that can detect six ROIs at a single step with high accuracy using a deep learning technique, YOLOv3[24], which is known for short execution time and high accuracy. Existing networks include a tremendous number of background classes, which are accompanied by hard negative mining and eventually slow the network down. YOLOv3 can ensure not only rapid performance rates through the use of full images without negative mining but also comparable high accuracy, especially for detecting small objects such as the metaphyseal bones of hand and wrist.

The accuracy of the system trained with only SMIs was 0.46 and 1.24 of MAE in males and females, respectively, which corresponds to 0.28 and 0.94 years when applying the mean age interval of the modified Korean CA-SMI chart as shown in Table 11. This was on par with measurements from prior studies on deep learning systems for BAA.[15,35]

However, it was unfavorable that the top-1 concordance rate in SMI 6, which is comparable to the pubertal growth spurt, showed a low value of 22.2 %. Moreover, the top-1 accuracy for SMI 9 and 10 of males and SMI 3-5 of females was 0.0 %. It is worthy of note that the top-1 accuracy was 100.0 % in males while 0.0 % in females for SMI 4, which is the stage that seems relatively easy to determine because it primarily relies on whether the adductor sesamoid of thumb appeared. Class imbalance by SMIs and smaller hand sizes of females could be the influencing factors to consider in this regard.

For the system trained with only SMIs, attention maps were extracted to identify important regions or structures for estimating skeletal maturity for each SMI (Figure 8). While approximately half of the test images were focused on the area corresponding to the resulted SMI, i.e., the regions taken into account by radiologists for the Fishman method, the others were done on the whole hand and wrist or only the very minor area such as a fingertip. There were even a few cases activated just over the background with no mapping on the hand-wrist area as well. While Lee et al.[14] achieved the expected mapping for the GP-based system, Souza et al.[36] reported that a CNN model treated the phalanges alone (not the wrist area) as the most important structures in the post-puberty stage. The fact that the regions paid attention to by the system do not match the regions considered by radiologists might suggest ways for understanding bone maturation and improving conventional BAA methods.

On the other hand, it might be natural that better accuracy was obtained in the system with ROI-based training, which could effectively concentrate on discriminative localized features. The MAE in SMI prediction was 0.34 and converted to 0.22 years based on the average CA interval between SMIs, which outperforms the values of reported existing deep learning systems for BAA so far. In addition, the average accuracy for ROI extraction and skeletal maturity determination for each ROI was very high. Compared to the other ROIs, ROI 4 showed the lowest concordance rate for skeletal maturity prediction for each ROI, correspondingly, it was the most difficult region for radiologists to evaluate due to the small size and the configurational variation of the epiphysis.

Benefiting from the potential of deep learning, a type of reverse-direct approach can be performed to train the systems using many images of heathy subjects labeled by CAs and find which regions are highlighted in attention maps regarding age. For this, it is inevitable

to improve the reliability of the mapping at an optimal level.

In the analysis and diagnosis of oral and maxillofacial images as well as medical images, the possibilities of deep learning applications are endless. Recently, a system for analyzing alveolar bone levels in dental panoramic radiographs was released.[37] Deep learning techniques could help to efficiently identify the presence and location of dental caries, periapical lesion, and diseases such as cyst, and benign and malignant tumors. Diagnosis using deep learning systems is expected to be corroborated by a doctor, even in quite a distant future, but it could eliminate the problem of varied interpretation quality depending on the experience or expertise of the given doctor and facilitate the diagnostic purpose of obtaining as much information as possible from images. Additionally, new radiological findings that have not been discovered by humans may be unexpectedly identified, benefiting from feature learning and excellent generalization ability of deep learning. For this, it is necessary to use a large amount of training data.

This study had several limitations. First, it may be biased given that all of the included hand-wrist radiographs were taken from just one clinical site, and the histogram equalization, a preprocessing technique, would contribute to data homogenization. Second, only a part of the sample was assessed by the TW3 method. The third was sample imbalance regarding SMI. The insufficient number of images for specific SMIs may be one of the factors that decreased the system accuracy. Lastly, the scheme that Fishman used to originally define the order of ROI observation[8] was not considered for developing the systems in this study. Instead, the ideal combination of the maturation stages by ROIs was set up for each SMI as shown in Table 4. Further research is needed to develop a system that can take into account Fishman's ROI observation order.

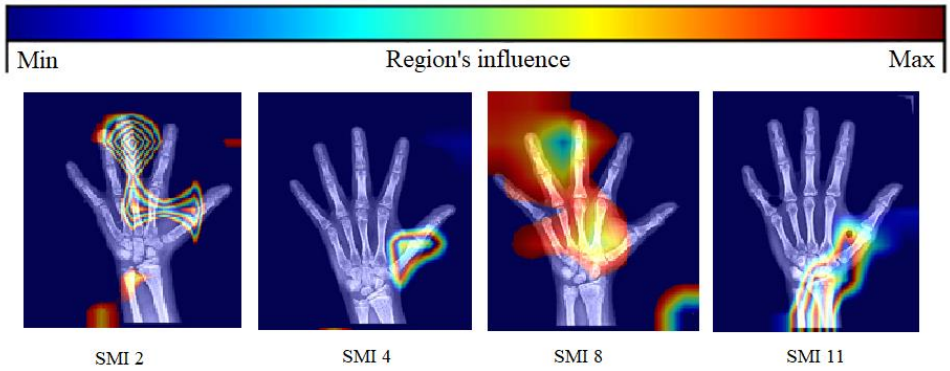


Figure 8. Examples of attention maps for the system trained with the dataset labeled with only skeletal maturity indicators (SMIs)

V. Conclusion

In conclusion, it was confirmed that Fishman's SMIs are a significantly reliable index for the BAA of contemporary Korean children and adolescents. The developed deep learning model could automate the entire process consisting of ROI extraction, skeletal maturity determination for each ROI, and final SMI prediction. The skeletal maturity prediction accuracy of the system outperformed that of existing fully-automated systems for BAA. The system in the study could be readily used for growth evaluation in clinical practice.

VI. References

1. Benjavongkulchai S, Pittayapat P. Age estimation methods using hand and wrist radiographs in a group of contemporary Thais. Forensic Sci Int 2018; 287: 218e1-8.
2. Mappes MS, Harris EF, Behrents RG. An example of regional variation in the tempos of tooth mineralization and hand-wrist ossification. Am J Orthod Dentofacial Orthop 1992; 101: 145-51.
3. Liversidge HM, Herdeg B, Rösing FW. Dental age estimation of non-adults: a review of methods and principles. In: Alt KW, Rösing FW, Teschler-Nicola M. Dental anthropology. Vienna: Springer; 1998. p. 419-42.
4. Demirjian A, Buschang PH, Tanguay R, Patterson DK. Interrelationships among measures of somatic, skeletal, dental, and sexual maturity. Am J Orthod 1985; 88: 433-8.
5. Spampinato C, Palazzo S, Giordano D, Aldinucci M, Leonardi R. Deep learning for automated skeletal bone age assessment in X-ray images. Med Image Anal 2017; 36: 41-51.
6. Gilsanz V, Ratib O. Hand bone age: A digital atlas of skeletal maturity. Berlin: Springer; 2005.
7. Tanner JM, Healy MJR, Cameron N, Goldstein H. Assessment of skeletal maturity and prediction of adult height (TW3 method). Philadelphia: W. B. Saunders; 2001.
8. Fishman LS. Radiographic evaluation of skeletal maturation. a clinically oriented methid based on hand-wrist films. Angle Orthod 1982; 52: 88-112.
9. Fishman LS. Maturational patterns and prediction during adolescence. Angle Orthod 1987; 57: 178-93.
10. Khan K, Elayappen AS. Bone growth estimation using radiology (Greulich–Pyle and Tanner–Whitehouse methods). In: Preedy VR. Handbook of growth and growth monitoring in health and disease. New York: Springer; 2012. p. 2937-53.
11. Flores-Mir C, Nebbe B, Major PW. Use of skeletal maturation based on hand-wrist radiographic analysis as a predictor of facial growth: a systematic review. Angle Orthod 2004; 74: 118-24.

12. Phulari BS. Orthodontics: principles and practice. London: Jaypee Brothers Medical Publishers; 2016.
13. Kim JR, Shim WH, Yoon HM, Hong SH, Lee JS, Cho YA, et al. Computerized bone age estimation using deep learning based program: evaluation of the accuracy and efficiency. AJR Am J Roentgenol 2017; 209: 1374-80.
14. Lee H, Tajmir S, Lee J, Zissen M, Yeshwas BA, Alkasb TK, et al. Fully automated deep learning system for bone age assessment. J Digit Imaging 2017; 30: 427-41.
15. Larson DB, Chen MC, Lungren MP, Halabi SS, Stence NV, Langlotz CP. Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs. Radiology 2018; 287: 313-22.
16. Son SJ, Song Y, Kim N, Do Y, Kwak N, Lee MS, et al. TW3-based fully automated bone age assessment system using deep neural networks. IEEE Access 2019; 7: 33346-58.
17. Chen M. Automated bone age classification with deep neural networks. Stanford University, Tech. Rep; 2016.
18. Thodberg HH, Kreiborg S, Juul A, Pedersen KD. The BoneXpert method for automated determination of skeletal maturity. IEEE Trans Med Imaging 2009; 28: 52-66.
19. Seok J, Hyun B, Kasa-Vubu J, Girard A. Automated classification system for bone age X-ray images. 2012 IEEE International Conference on Systems, Man, and Cybernetics. 2012: 208-13.
20. Iglovikov VI, Rakhlin A, Kalinin AA, Shvets AA. Paediatric bone age assessment using deep convolutional neural networks. In: Stoyanov D, Taylor Z, Carneiro G, Syeda-Mahmood T, Martel A, Maier-Hein L, et al. Deep learning in medical image analysis and multimodal learning for clinical decision support. Cham: Springer; 2018. p. 300-8.
21. Haralick RM, Sternberg SR, Zhuang X. Image analysis using mathematical morphology. IEEE Trans Pattern Anal Mach Intell 1987; 9: 532-50.
22. Hum YC, Lai KW, Salim MIM. Multiobjectives bihistogram equalization for image contrast enhancement. Complexity 2014; 20: 22-36.
23. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. 2016

IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-78.

24. Redmon J, Farhadi A. YOLOv3: An Incremental Improvement 2018; arXiv e-print:[arXiv:1804.02767p.]. Available from: https://arxiv.org/abs/1804.02767.
25. Szegedy C, Toshev A, Erhan D. Deep neural networks for object detection. Proceedings of the 26th international conference on neural information processing systems. 2013: 2553-61.
26. Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 580-7.
27. Mughal AM, Hassan N, Ahmed A. The applicability of the Greulich & Pyle Atlas for bone age assessment in primary school-going children of Karachi, Pakistan. Pak J Med Sci 2014; 30: 409-12.
28. Cantekin K, Celikoglu M, Miloglu O, Dane A, Erdem A. Bone age assessment: the applicability of the Greulich-Pyle method in eastern Turkish children. J Forensic Sci 2012; 57: 679-82.
29. Mohammed RB, Kalyan VS, Tircouveluri S, Vegesna GC, Chirla A, Varma DM. The reliability of Fishman method of skeletal maturation for age estimation in children of South Indian population. J Nat Sci Biol Med 2014; 5: 297-302.
30. Paxton ML, Lamont AC, Stillwell AP. The reliability of the Greulich-Pyle method in bone age determination among Australian children. J Med Imaging Radiat Oncol 2013; 57: 21-4.
31. Ontell FK, Ivanovic M, Ablin DS, Barlow TW. Bone age in children of diverse ethnicity. AJR Am J Roentgenol 1996; 167: 1395-8.
32. Hägg U, Taranger J. Maturation indicators and the pubertal growth spurt. Am J Orthod 1982; 82: 299-309.
33. Björk A, Helm S. Prediction of the age of maximum puberal growth in body height. Angle Orthod 1967; 37: 134-43.
34. Al-Jewair TS, Preston CB, Flores-Mir C, Ziarnowski P. Correlation between craniofacial growth and upper and lower body heights in subjects with Class I occlusion. Dental Press J Orthod 2018; 23: 37-45.
35. Rijn RR, Thodberg HH. Bone age assessment: automated techniques coming of age?

Acta Radiol 2013; 54: 1024-9.

36. Souza D, Oliveria MM. End-to-end bone age assessment with residual learning. 2018 31st SIBGRAPI Conference on Graphics, Patterns and Images. 2018: 197-203.
37. Chang HJ, Lee SJ, Yong TH, Shin NY, Jang BG, Kim JE, et al. Deep learning hybrid method to automatically diagnose periodontal bone loss and stage periodontitis. Sci Rep 2020; 10: 7531.

국문요약

현대 한국 소아 및 청소년의 수완부방사선영상을 이용한 골령 평가법 비교 및 Fishman법 기반의 골격 성숙도 측정 딥러닝 시스템 개발

신 난 영

서울대학교 대학원
치의과학과 영상치의학 전공
(지도교수 허 민 석)

목적

수완부방사선영상을 이용한 골령 평가에는 Greulich-Pyle(GP)법, Tanner-Whitehous 3(TW3)법 및 Fishman법이 주로 이용되고 있다. 본 연구에서는 건강한 현대 한국 소아 및 청소년의 수완부방사선영상을 이용해 위의 3가지 골령 평가 방법을 비교하여 Fishman의 골격성숙도지수(skeletal maturity indicator; SMI)의 유용성을 검증하고, SMI를 기반으로 하여 심층 신경망을 이용한 골격 성숙도 측정 딥러닝 시스템을 개발한 후 시스템의 정확도를 평가하고자 하였다.

재료 및 방법

본 연구를 위해 2012-2017년에 촬영된 좌측 수완부방사선영상 1,617매(남: 706, 여: 911; 연령 6-17세)를 수집하였다. GP법, TW3법 및 Fishman법에 따라 측정한 골령과 연대기적 연령의 관계를 대응 표본 t 검정 및 상관관계 분석을 이용하여 비교 평가하였다. 또한 Fishman법에 기반하여 골격 성숙도를 측정하는 완전 자동화 딥러닝 시스템을 개발하고자, SMI만으로 라벨링한 데이터를 이용해 학습을 시행한 시스템과 관심 부위, 관심 부위 별 골격 성숙도 및 SMI로 라벨링한 데이터를 이용해 학습을 시행한 시스템을 각각 개발하여 이들의 정확도를 평가하였다. SMI에 대한 참조 표준은 영상치의학 판독의 2인이 토의하여 작성하였다.

결과

TW3법의 남성 그룹을 제외하고는 3가지 방법에 있어 전체 그룹 및 성별 하위 그룹에서 골령과 연대기적 연령 사이에 유의한 차이가 있었다. 그러나 3가지 방법에서 모두 골령과 연대기적 연령 사이에 높은 상관관계가 관찰되었다. SMI만으로 라벨링한 데이터를 이용해 학습을 시행한 시스템의 경우, SMI의 평균 절대 오차 값이 0.88, 1등급 내 일치도는 73.1 %였다. 관심 부위 추출, 관심 부위 별 성숙도 평가 및 최종 SMI 평가로 구성된 시스템의 경우, 보다 우수한 결과를 나타내어 그 값이 각각 0.34, 93.7 %였다.

결론

본 연구에서 Fishman의 SMI는 수완부방사선영상을 이용한 골격 성숙도 측정에 있어 신뢰할 만한 척도로 평가되었다. 또한 본 연구에서 관심 부위 추출, 관심 부위 별 골격 성숙도 측정 및 최종 SMI 예측으로 구성되는 전 과정을 자동화한 딥러닝 시스템이 개발되었으며, 이는 골격 성숙도 측정에 있어 우수

한 정확도를 보여주었다. 따라서 개발된 딥러닝 시스템은 현대 한국 아동 및 청소년의 골격 성숙도 측정에 유효하게 사용될 수 있을 것이다.

주요어: 골격 성숙도 지수, 수완부방사선사진, Fishman법, 딥러닝
학번: 2017-35685